What Supercomputers Say: A Study of Five System Logs

Adam J. Oliner

Stanford University

Jon Stearley

Sandia National Laboratories

DSN, June 26th, 2007

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy's National Nuclear Security Administration
under contract DE-AC04-94AL85000.

Today’s Menu

- Motivation
- Data
- Seven Insights
- Recommendations

The Goal

- **Use system logs to**
  - Detect faults
  - Attribute root causes
  - Predict failures
  - Quantify RAS
- **NOT to compare systems**
  - “absurd”

Why System Logs?

Log Message Examples

- NULL RAS BGLMASTER FAILURE ciodb exited normally with exit code 0

- kernel: VIPKL(1): [create_mr] MM_bld_hh_mr failed (-253:VAPI_EAure = no

- kernel: Uhhuh. NMI received. Dazed and confused, but trying to continue

- kernel: Losing some ticks… checking if CPU frequency changed.

The Systems

SYSTEM	RANK	PROCS	MEMORY (GB)	DURATION (Days)
Blue Gene/L	1	131072	32768	215
Thunderbird	6	9024	27072	244
Red Storm	9	10880	32640	104
Spirit	202	1028	1024	558
Liberty	445	512	944	315

Alerts

- **Alert**
  - Message of interest to system administrators
- **Failure**
  - Event of interest
  - Mapping is many-to-many

Alert Tagging

- **Combination of rules and manual labor**
- **178,081,459 alerts**
- **Severity field**
  - 59% false positive rate (BG/L)
  - Often unrecorded (Thunderbird, Spirit, Liberty)

Our Distinctions

- **Largest system log study to date**
  - 111.67 GB
  - ~1 billion messages
  - 774 million processor hours
- **Raw logs from five supercomputers**
- **Manual alert tagging**

Prior Work

- **Derived data**
  - [Schroeder, 06]
- **Simplistic tagging strategies**
- **Small systems**
- **System-specific**
  - [Liang, 06]
- **Models of convenience**

Seven Insights

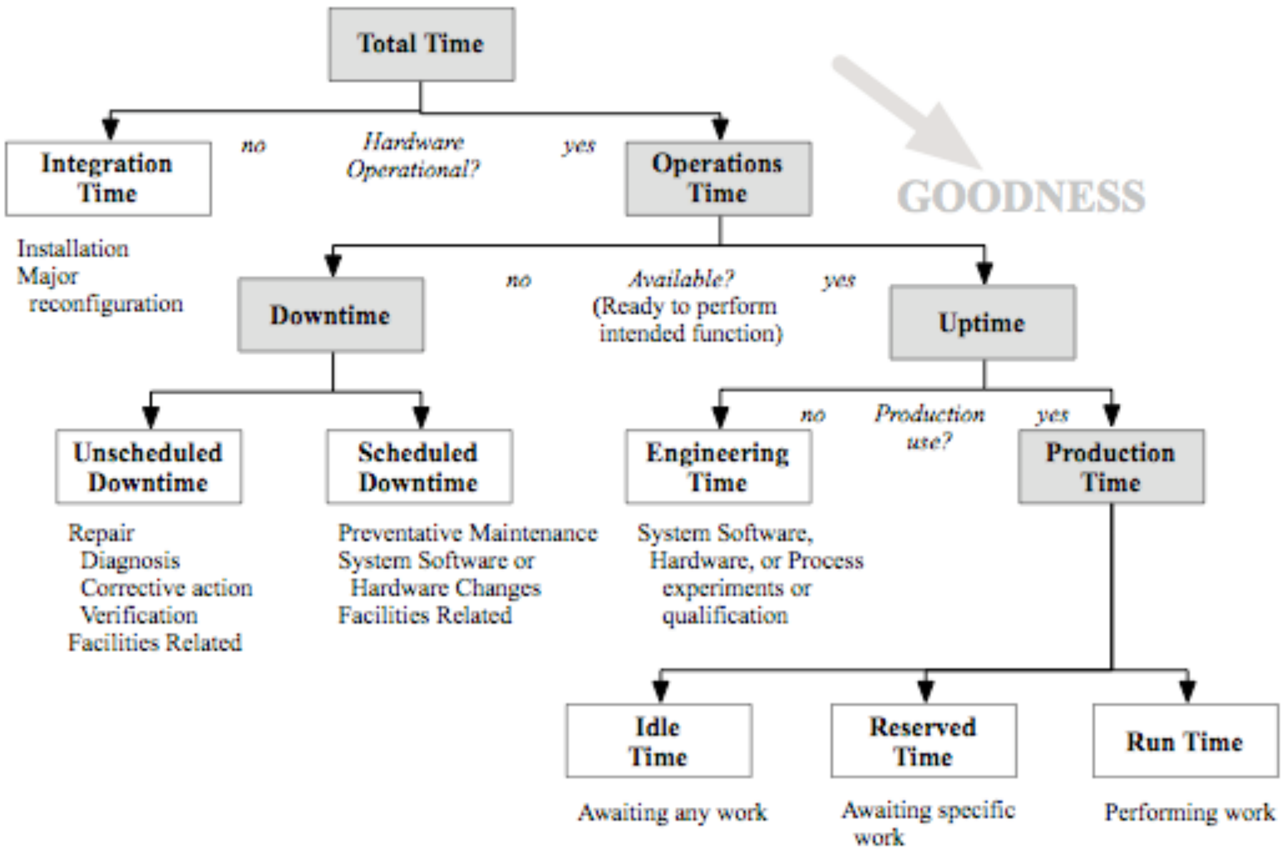
1. **Insufficient Context**
2. **System Evolution**
3. **Implicit Correlation**
4. **Inconsistent Structure**
5. **Corruption**
6. **Redundancy**
7. **Misdirection**

1. Insufficient Context

-
  ```
  NULL RAS BGLMASTER FAILURE ciodb
  exited normally with exit code 0
  ```
- **Two meanings:**
  - Everything is fine
  - Every job died
- **Requires operational context**

1. Insufficient Context

State Diagram



Total Time

Hardware Operational?

no

Integration Time

Installation
Major
reconfiguration

yes

Operations Time

GOODNESS

Available?
(Ready to perform intended function)

no

Downtime

Unscheduled Downtime

Repair
Diagnosis
Corrective action
Verification
Facilities Related

Scheduled Downtime

Preventative Maintenance
System Software or
Hardware Changes
Facilities Related

yes

Uptime

Production use?

no

Engineering Time

System Software,
Hardware, or Process
experiments or
qualification

yes

Production Time

Idle Time

Awaiting any work

Reserved Time

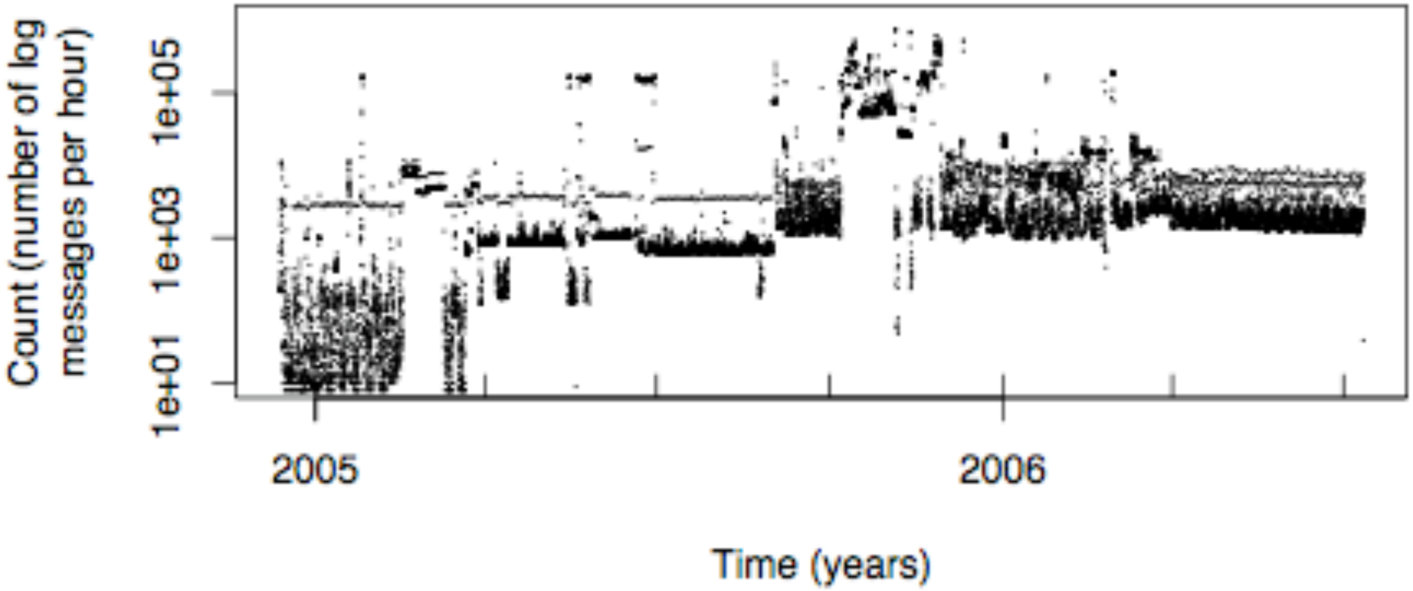
Awaiting specific work

Run Time

Performing work

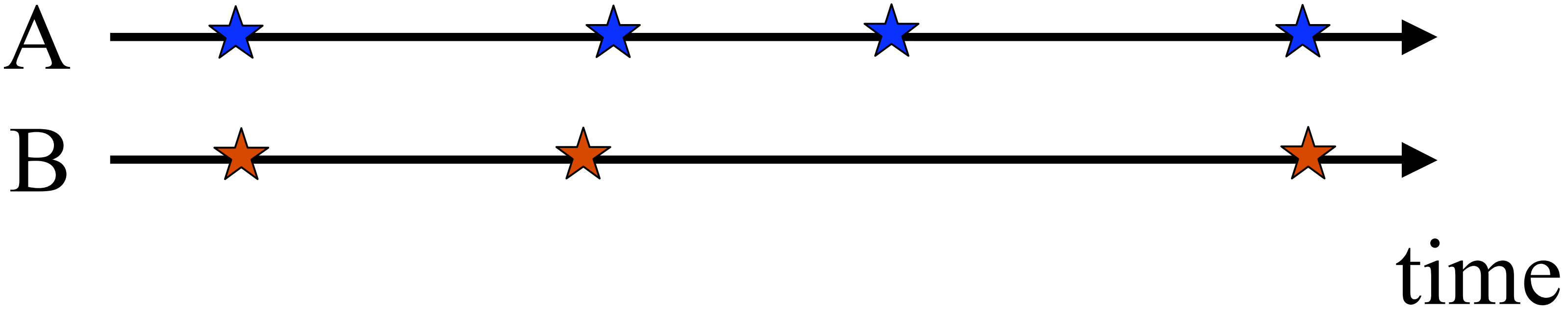
2. System Evolution

- Moving target
- Need to detect phase shifts



3. Implicit Correlation

- **Messages may be related**
  - Similar code paths
  - Similar triggers



4. Inconsistent Structure

- [YYYY-OO-DD-HH.MM.SS.UUUUUU] [rack]-[midplane]-[node]-[core] [subsystem] [sender] [severity] [message body]

- [YYYY-OO-DD] [HH:MM:SS]|[YYYY-OO-DD] [HH:MM:SS]|[file source]|src:::[source id]|svc:::[svc id]|[message body]

- [Facility/Severity Hex] [Month] [DD] [HH:MM:SS] [source] [message body]

- [Month] [DD] [HH:MM:SS] [source1]/[source2] [message body]

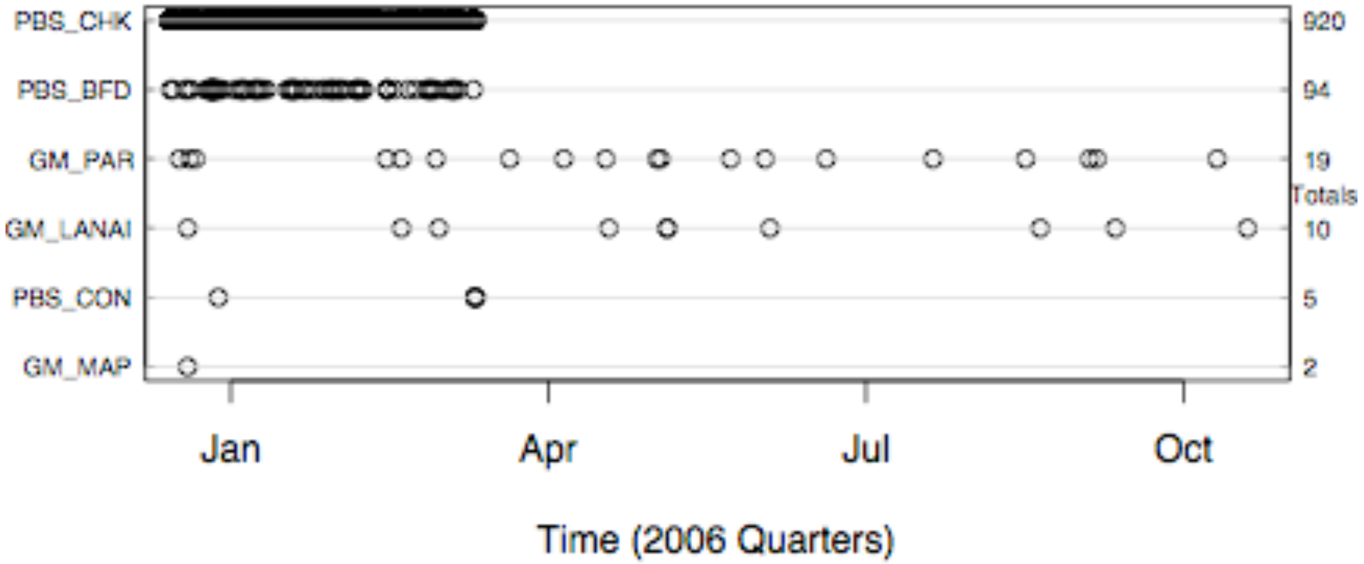
5. Corruption

```
•kernel:  VIPKL(1): [create_mr] MM_bld_hh_mr
 failed (-253:VAPI_EAGAIN)


•kernel:  VIPKL(1): [create_mr] MM_bld_hh_mr
 failed (-253:VAPI_EAure = no
•kernel:  VIPKL(1): [create_mr] MM_bld_hh_mr
 failed (-253:VAPI_EAGsys/mosal_iobuf.c
 [126]: dump iobuf at 0000010188ee7880 :
•kernel:  VIPKL(1): [create_mr] MM_bld_hh_mr
 failed (-253:VAPI_EAGAI
```

6. Redundancy

- **In six days on Spirit**
  - One disk problem
  - 56,793,797 alerts
- **But,**



7. Misdirection

- `kernel: Losing some ticks… checking if CPU frequency changed.`
- What does this mean?
- Hint: Correlated across nodes!
- Answer: Bug in OS; missed interrupts under heavy network activity.

Recommendations

- Avoid severity field
- Log operational context
- Be aware of the insights
- Measure metrics of interest directly

… One More Thing

- **We are please to announce the public availability of these logs, starting today**
  - Some scrubbing of sensitive data
  - Initially by request

oliner@cs.stanford.edu

jrstear@sandia.gov